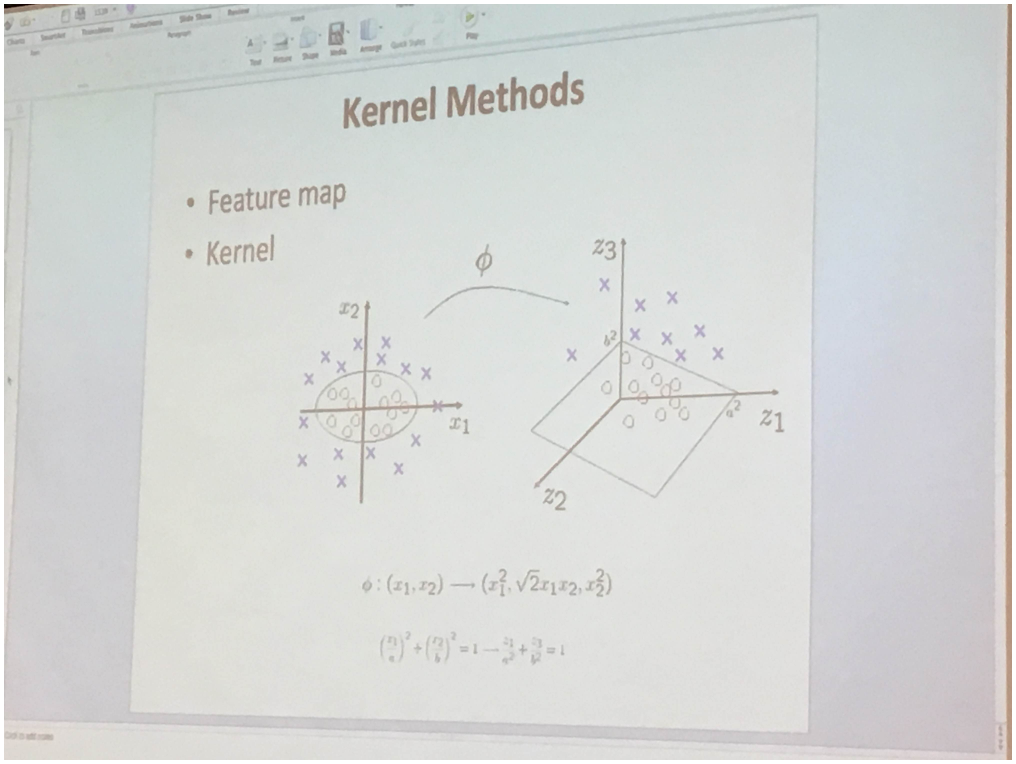
Mathematics of Big Data I

Lecture 6: SVM (Conclusion), Lagrange Duality, KKT Conditions, Kernel Methods

October 10th, 2016

1 Introduction



The kernel method is related to the *dual view* of the support vector machine. Suppose for instance we have an ellipse in the plane:

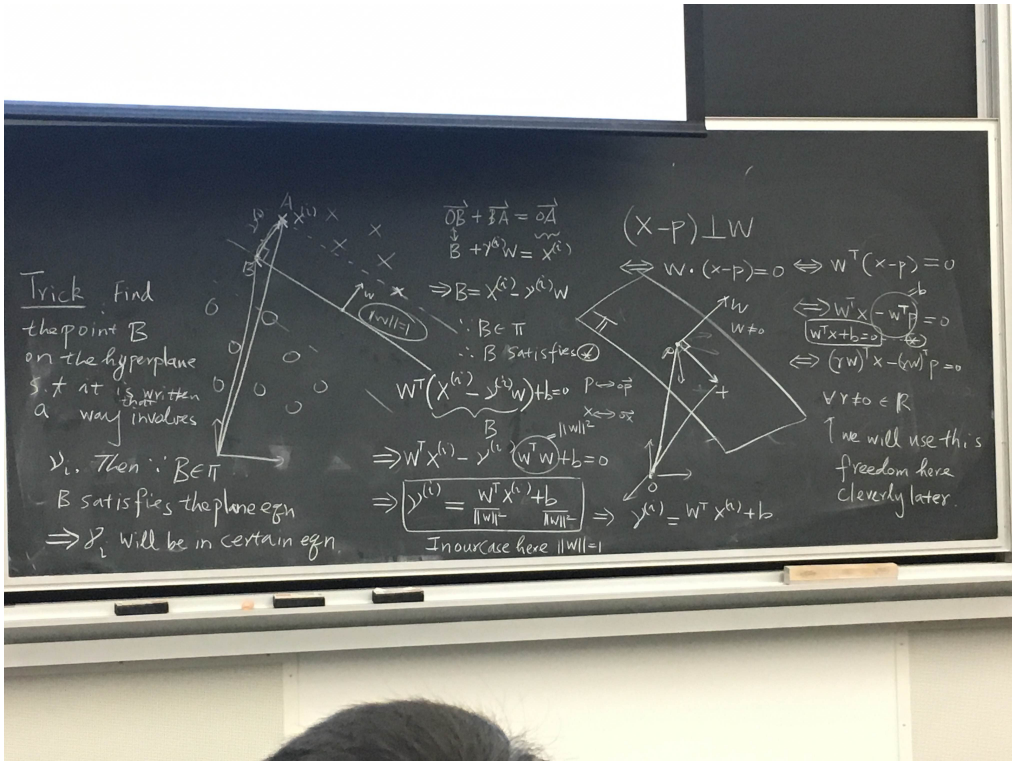
$$\left(\frac{x_1}{a}\right)^2 + \left(\frac{x_2}{b}\right)^2 = 1.$$

We can define the map $\phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ and then eventually we will want to compute our solution in terms of the inner product $\langle \phi(x), \phi(y) \rangle$ and therefore we can transform a nonlinear problem into a linear one.

2 Support Vector Machine (Continued)

Recall that the SVM method is based on a geometrically intuitive argument so that we can draw a decision boundary between two different classes of data. Today our focus will be on the dual method, so that we

can express everything in terms of an inner product. Remember that the functional margin is given by $\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$ with respec to a certain training sample and we can take $\hat{\gamma} = \min_i \hat{\gamma}^{(i)}$.



The method from SVM can be extended to any quantity we're interested in. For example, suppose we want to find the point B on the hyperplane such that it is written in a way that involves $\gamma^{(i)}$. Therefore since B is in the hyperplane B satisfies the plane equation and therefore $\gamma^{(i)}$ will satisfy a certain equation. For example, if x and p are two points on the plane, we want our support vector w to satisfy $(x-p) \perp w$ if and only if $w^T(x-p) = 0$ if and only if $w^T x + b = 0$ where we take $b = -w^T p$, and therefore for all nonzero $r \in \mathbb{R}$ we can write $(rw)^T x - (rw)^T p = 0$. We will take advantage of this freedom with r later. If we divide the functional margin equation by the norm of w squared, we obtain the **geometric margin**:

$$\gamma^{(i)} = \frac{w^T x^{(i)}}{\|w\|^2} + \frac{b}{\|w\|^2}.$$

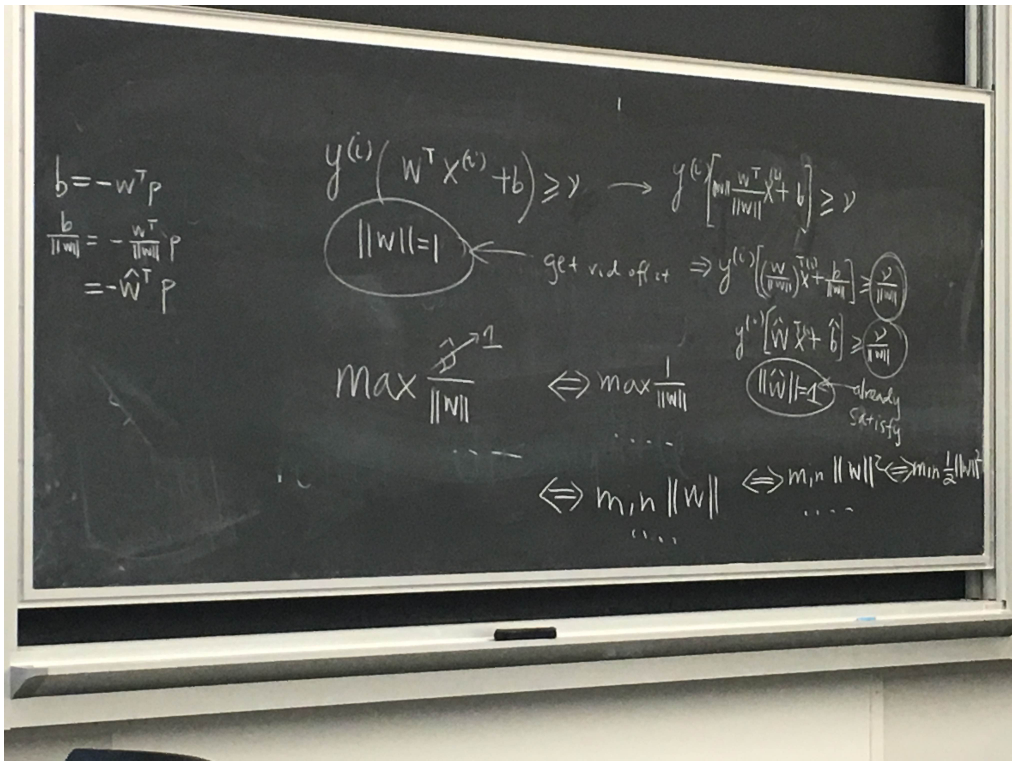
The problem here is that this is a nonconvex problem. We can however transform this into a convex problem. We can inherently build in $\|w\| = 1$ and utilize this to obtain a convex problem. We can instead turn our optimization problem into maximizing γ subject to the constraints that $\|w\| = 1$ and

$$y^{(i)}(w^T x^{(i)} + b) \geq \gamma$$

for $i \in \{1, \ldots, m\}$. Starting with this last conditions we can cleverly write this as

$$\begin{aligned} y^{(i)}(w^T x^{(i)} + b) &= y^{(i)}\left(\frac{w}{\|w\|}\right)^T x^{(i)} + \frac{b}{\|w\|} \geq \frac{\gamma}{\|w\|} \\ &= y^{(i)}\hat{w}^T x^{(i)} + \hat{b} \geq \frac{\gamma}{\|w\|}. \end{aligned}$$

Our optimization now becomes to maximize $\frac{\hat{\gamma}}{\|w\|}$ which occurs if and only if we maximize $\frac{1}{\|w\|}$ which occurs if and only if we minimize $\|w\|$ which occurs if and only if we minimize $\frac{1}{2}\|w\|^2$, and therefore we have a convex problem.



3 Lagrange Duality

We want to take what we learned in constructing the SVM to developing a new form of machine learning called **kernel methods**. First we need to look at the SVM from a *dual* point of view, so that we can express everything in terms of an inner product. Consider the general problem of

$$\min_w f(w) \qquad \text{such that} \qquad h_i(w) = 0$$

for $i \in \{1, \ldots, l\}$. We define the **Lagrangian** as

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

The components of $\beta = \begin{bmatrix} \beta_1 & \ldots & \beta_l \end{bmatrix}^T$ are called the **Lagrange multipliers**. In a traditional Lagrange multiplier problem we take the partial derivatives $\frac{\partial \mathcal{L}}{\partial w_i} = 0$ and $\frac{\partial \mathcal{L}}{\partial \beta_j} = 0$ and solve for w and β in these situations. Now this problem is easy if these equations are constrained to be equal, but in general we have to deal with *inequalities*. The problem here is we want $\min_w f(w)$ subject to the constraints there are functions $g_i(w) \leq 0$ and $h_j(w) = 0$ for $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, l\}$. This kind of optimization is called a **primal optimization problem**. The **generalized Lagrangian** is given by

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{j=1}^{l} \beta_j h_j(w).$$

How do we get rid of the inequalities in our constraints? Consider the quantity

$$\theta_{\mathcal{P}}(w) = \max_{\alpha, \beta; \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

where the $\mathcal{P}$ denotes the primal optimization. To start this, suppose some quantity w is chosen. If w violates any of the primal constraints then you can verify that $\theta_{\mathcal{P}}(w) = \infty$. Conversely if the constraints are required to satisfy $\theta_{\mathcal{P}} = f(w)$ for a particular value of w then we can conclude that

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w), & \text{if } w \text{ satisfies the primal constraints} \\ \infty, & \text{otherwise} \end{cases}.$$

We can therefore get rid of the inequalities since our original problem is equivalent to

$$\min_w f(w) = \min_w \theta_{\mathcal{P}}(w) = \min_w \max_{\alpha,\beta;\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta).$$

For later use we also define the optimal value of the objective $p^* = \min_w \theta_{\mathcal{P}}(w)$. If we consider a different problem, say

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_w \mathcal{L}(w, \alpha, \beta)$$

then we can solve the **dual problem**. The difference here is we find that minimum of the Lagrangian with respect to w and then follow this by maximizing with respect to the coordinates α, β. The dual problem thus stated is given by

$$\max_{\alpha,\beta;\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta;\alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta).$$

The obvious question to ask here is **does the solution to the dual problem coincide with that of the primal problem?** In general the answer is no. If the solution of the dual problem is given by d^*, then we have that $d^* \leq p^*$ for all Lagrangians. Under certain conditions we do obtain equality, which is important since the dual problem can be formulated strictly in terms of an inner product. What are these conditions?

1. The functions f, g_i are all convex functions and the h_j are affine.
2. The constraints g_j are *strictly feasible* meaning there exists some w such that $g_j(w) < 0$ for all j.

These conditions yield the existence of some w^*, α^*, β^* so that w^* is the solution to the primal problem and α^*, β^* are the solutions to the dual problem. Moreover $p^* = d^*$ and $\mathcal{L}(w^*, \alpha^*, \beta^*)$ satisfy the **Kuhn-Kusher-Tucker or KKT Conditions**:

$$\begin{aligned} \frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0 \\ \frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0 \\ \alpha_i^* g_i(w^*) &= 0 \\ g_i(w^*) &\leq 0 \\ \alpha^* &\geq 0. \end{aligned}$$

The **dual complimentary condition** is that $\alpha_i^* g_i(w^*) = 0$ for all $i \in \{1, \ldots, k\}$. This implies that if $\alpha_i^* > 0$ the $g_i(w^*) = 0$. This will later be key in showing that SVM needs only a small number of support vectors, and will also be of importance in showing convergence of the SMO algorithm.

Recalling the convex form of SVM, our problem is

$$\min_{\gamma,w,b} \frac{1}{2}\|w\|^2 \qquad \text{such that} \qquad y^{(i)}(w^T x^{(i)} + b) \geq 1.$$

We can simply take

$$g_i = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0.$$

To find the dual problem we need to minimize $\mathcal{L}$ with respect to w and b first for fixed α to get $\theta_{\mathcal{D}}$. We therefore have

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m}\alpha_i\left[y^{(i)}(w^T x^{(i)}+b)-1\right]$$

and to minimize we get

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)} = 0.$$

We eventually arrive at the dual optimization problem:

$$\begin{aligned}\max_{\alpha} W(\alpha) &= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle \\ &\text{such that} \qquad \alpha_i \geq 0 \qquad \text{and} \qquad \sum_{i=1}^{m}\alpha_i y^{(i)} = 0.\end{aligned}$$

The cool thing about this problem is that we can rewrite the quantity

$$\begin{aligned}w^T x + b &= \left(\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)}\right)^T x + b \\ &= \sum_{i=1}^{m}\alpha_i y^{(i)}\left\langle x^{(i)}, x\right\rangle + b.\end{aligned}$$

4 Kernel Methods

Duality plays a very important role in solving a certain class of nonlinear problems. This falls in the realm of **kernel methods**. We want to utilize something close to an inner product, but with an important modification. To illustrate what we're interested in, consider $K(x,z) = (x^T z)^2$ where $x, z \in \mathbb{R}^n$. After some computation we can write

$$K(x,z) = \sum_{i,j}^{n}(x_i x_j)(z_i z_j).$$

Let's suppose instead that we write the function

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}.$$

then we can write $K(x,z) = \phi(x)^T\phi(z)$, hence we can write K in terms of an inner product of the function of it's coordinates. More broadly the **kernel** $K(x,z) = (x^T z + c)^d$ corresponds to a feature mapping to an $\binom{n+d}{d}$ dimensional feature space corresponding to all monomials of the form $x_{i_1}x_{i_2}\ldots x_{i_k}$ that are up to order d. To formalize this

Definition 1 *Given a feature mapping ϕ we define the corresponding **kernel** as the mapping $K(x,z) = \phi(x)^T\phi(z) = \langle\phi(x),\phi(z)\rangle$.*

The kernel can be viewed as a way to measure similarities. For instance by measure the inner product $\phi(x)^T\phi(z)$ we can infer how similar the values x and z are. An example of this is the Gaussian kernel:

$$K(x,y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right).$$

This can be a reasonable measure since x and y will be very close to each other if $K(x,y) \approx 1$ and they are very far if $K(x,y) \approx 0$. The feature mapping here is actually infinite-dimensional. More broadly though we ask if given a kernel K is there some feature mapping ϕ such that $K(x,z) = \langle\phi(x),\phi(y)\rangle$. To answer this question we introduce the **kernel matrix** K where given a training set $\{x^{(1)},\ldots,x^{(m)}\}$ then the matrix K has entries $K_{ij} = \langle x^{(i)}, x^{(j)}\rangle$. It is straight-forward to show that K is symmetric and positive semi-definite, meaning that $z^TKz \geq 0$ for all z.

Theorem 1 (Mercer) *Let $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be given. Then for K to be a valid kernel it is necessary and sufficient that for any $\{x^{(1)},\ldots,x^{(m)}\}$ the corresponding kernel matrix is symmetric positive semi-definite.*